Accuracy of the Perceptual Evaluation of Speech Quality (PESQ) algorithm

Scott Pennock
Lucent Technologies
Fax: 630.857.2376
E-mail:spennock@lucent.com

ABSTRACT

The Perceptual Evaluation of Speech Quality (PESQ) algorithm is the current industry standard for objective prediction of one-way speech quality. It is argued that the available performance data present an overly optimistic view of PESQ accuracy. Accuracy experienced by the telecommunications industry in real-world applications is likely to be worse. The current paper presents results from an investigation of PESQ accuracy. The effects of speech coder, packet loss concealment strategy, IP payload size, packet loss rate, and "burstiness" of packet losses on PESQ accuracy were assessed. The results indicate that the PESQ algorithm is a useful tool in helping identify potential performance problems. However, there are limitations to using PESQ for verification of voice quality performance, competitive analysis, and system optimization. It is concluded that important decisions should be based on the results from perceptual studies, and that PESQ is not accurate enough to specify speech quality requirements in Service Level Agreements (SLAs).

Keywords: objective quality; speech quality; MOS

INTRODUCTION

Currently, the most accurate way to assess speech quality of next generation communications networks (e.g., wireless, VoIP, etc.) is through subjective testing (i.e., perceptual testing). However, subjective testing has some disadvantages. It is time-consuming, expensive, and requires a lot of resources. Furthermore, monitoring real-time performance is not practical.

Because of the disadvantages of subjective testing, there is a strong desire in the telecommunications industry to create a device capable of predicting speech quality from objective (i.e., physical) measurements of the communications network. Unfortunately, the task of quantifying speech quality using objective measurements is non-trivial. The problem is that what is being measured is a perceptual attribute—not a physically measurable entity such as a signal's intensity. Therefore, transformations of physical level measures into perceptually relevant quantities need to be obtained. This requires accurate modeling of the human auditory system, which includes filtering by the peripheral auditory system, low-level neural processing, and higher-level cognitive processing. These effects need to be accounted for by objective measures of speech quality. Until they are, properly designed subjective testing will continue to be the most accurate way of assessing speech quality. None of the current objective predictors of speech quality, including PESQ, accurately models these perceptual and cognitive processes.

Of course, accurately modeling the human auditory system is not required to develop objective measurements that correlate fairly well with perceived quality. For example, on traditional telecommunications systems the Speech-to-Noise Ratio has provided reasonable estimates of speech quality even though calculating this measure does not attempt to directly model human perception. However, the problem with measures that do not accurately model human perception is that when the nature of impairments of the network change, then prediction of speech quality falls apart.

The ITU-T has recently gone through a competition to find the state-of-the-art in terms of objective prediction of speech quality. It is intended that this objective method be used by the telecommunications industry to measure perceived quality of network connections.

In this competition, the Perceptual Evaluation of Speech Quality (PESQ) algorithm was shown to

significantly outperform other objective speech quality models in the competition. In February 2001, PESQ was approved as ITU-T Recommendation P.862, and the previously recommended method (ITU-T Recommendation P.861) was deleted.

While PESQ represents the state-of-the-art in terms of objective prediction of perceived quality, it does not always accurately predict perceived quality. Performance data presented in ITU-T Recommendation P.862 presents a very optimistic view of PESQ accuracy that can be expected by the telecommunications industry. The current paper presents a more conservative view of PESQ accuracy.

The next section briefly describes the PESQ algorithm and it's intended use. Then, there is a discussion of how accuracy of PESQ is assessed. Next, performance data from the ITU-T is presented. This is followed by a presentation of performance data from Lucent subjective studies. Results are grouped in subsections that reflect the different ways the PESQ algorithm is likely to be used by the telecommunications industry. Finally, some conclusions regarding the accuracy and uses of the PESQ algorithm are made.

PERCEPTUAL EVALUATION OF SPEECH QUALITY (PESQ)

The Perceptual Evaluation of Speech Quality (PESQ) algorithm is an objective method of measuring speech quality. A detailed description of PESQ and its intended uses can be found in ITU-T Recommendation P.862 [1]. Basically, PESQ predicts subjective MOS scores by comparing speech recordings that have been transmitted through the network under test (i.e., "processed" speech files) with the original versions of these speech recordings that were input to the network under test (i.e., "reference" speech files).

It is important to note that PESQ only measures one aspect of transmission quality. ITU-T Recommendation P.862 puts it this way:

"*It should also be noted that the PESQ algorithm does not provide a comprehensive evaluation of transmission quality. It only measures the effects of one-way speech distortion and noise on speech quality. The effects of loudness loss, delay, sidetone, echo, and other impairments related to two-way interaction (e.g., center clipper) are not reflected in the PESQ scores. Therefore, it is possible to have high PESQ scores, yet poor quality of the connection overall*."

MEASURING THE PERFORMANCE PESQ

PESQ was developed to predict the Mean Opinion Scores (MOS) of end-to-end network quality as judged by a panel of listeners. Each listener rates "The quality of the connection" by selecting one of five options: "Bad", "Poor", "Fair", "Good", and "Excellent". Numbers are then assigned to these labels (1, 2, 3, 4, and 5 respectively) and the average of the numbers is taken to represent the Mean Opinion Score (MOS). For example, a system with a MOS of "3.2" would be considered "fair".

Since PESQ is supposed to measure MOS, accuracy of the PESQ algorithm must be determined by comparing PESQ with MOS scores. For example, suppose the MOS for a particular network connection was "3.2". If the PESQ algorithm is accurate, then the PESQ score will be very close to "3.2". On the other hand, PESQ scores that are not close to "3.2" indicate the PESQ algorithm is not very accurate. By comparing predicted and observed MOS scores across many network connections, accuracy of the PESQ algorithm can be assessed.

However, an issue with using MOS scores to assess PESQ accuracy is the fact that the MOS for a particular network connection can vary depending on the details of the subjective study. The participants in the study, the quality of the other network connections in the study, and the test laboratory used have all been shown to influence MOS scores. Therefore, some of the inaccuracies in predicting MOS scores may be due to the faults of MOS testing methodology—not problems with the PESQ algorithm.

In an attempt to deal with this problem, researchers have performed regression mapping on a per study basis before calculating measures of PESQ accuracy. For each subjective study, a 3rd order monotonically constrained polynomial was used to minimize the differences between PESQ and MOS scores. While this data fitting procedure will help reduce prediction errors that the PESQ algorithm should not be penalized for, it also minimizes prediction errors that PESQ should be penalized for. This causes measures of PESQ accuracy to appear better than they really are. The following example may help make this more clear.

In the current Lucent study both random and "bursty" packet losses were studied. Figure 1 shows PESQ scores for random and bursty packet losses for one speech coder at each of the packet loss rates. It can be seen that PESQ scores were similar for both random and bursty packet losses. PESQ treated degradation due to packet losses the same regardless of whether they were bursty or not. Figure 2 shows MOS scores for the same data. Here, it can be seen that MOS scores for the two types of loss are very different, and that the scores drop faster for bursty than random as packet loss rate increases. Results from other studies presented to the ITU-T from Nortel [2] and AT&T [3] have also shown similar results. Therefore, this is a clear instance where the difference between MOS and PESQ scores is caused by an inability of PESQ to accurately model human perception—not an unwanted error introduced by the design of this particular study. If regression mapping were performed on this study's data, it would greatly reduce the prediction errors that PESQ should be penalized for, and result in overly optimistic estimates of PESQ accuracy.

The Lucent test results presented here do not use regression mapping before calculating performance data. However, this approach provides a more realistic picture of PESQ performance, as experienced by the telecommunications industry, than if regression mapping were performed.

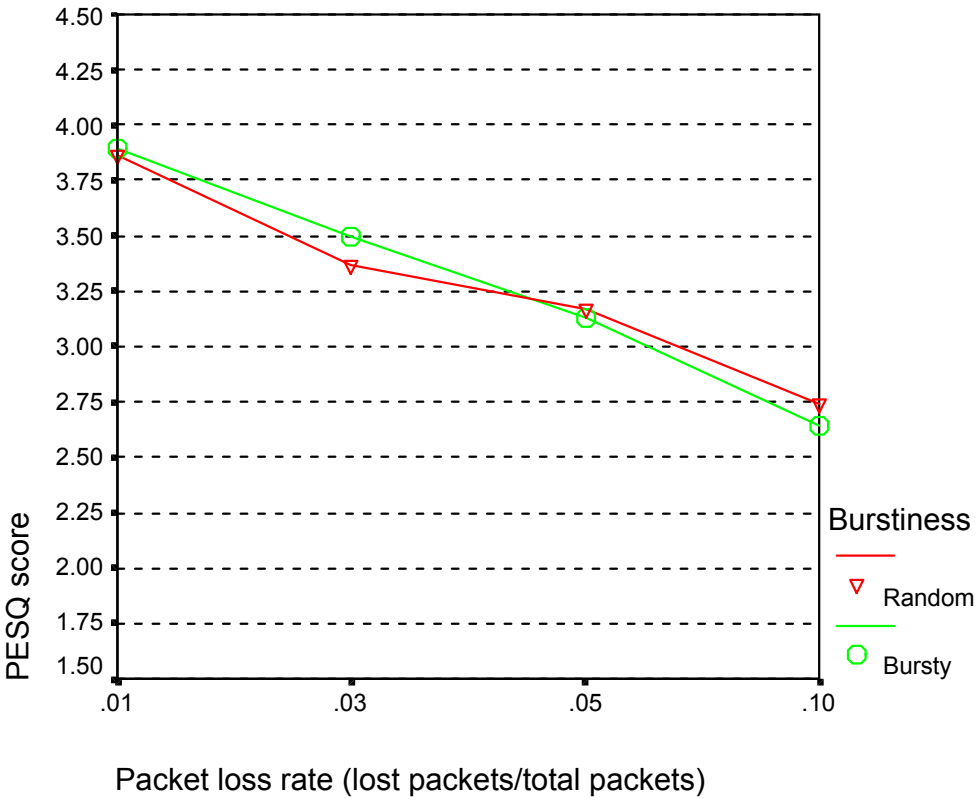

Figure 1. PESQ scores by "burstiness" and packet loss rate

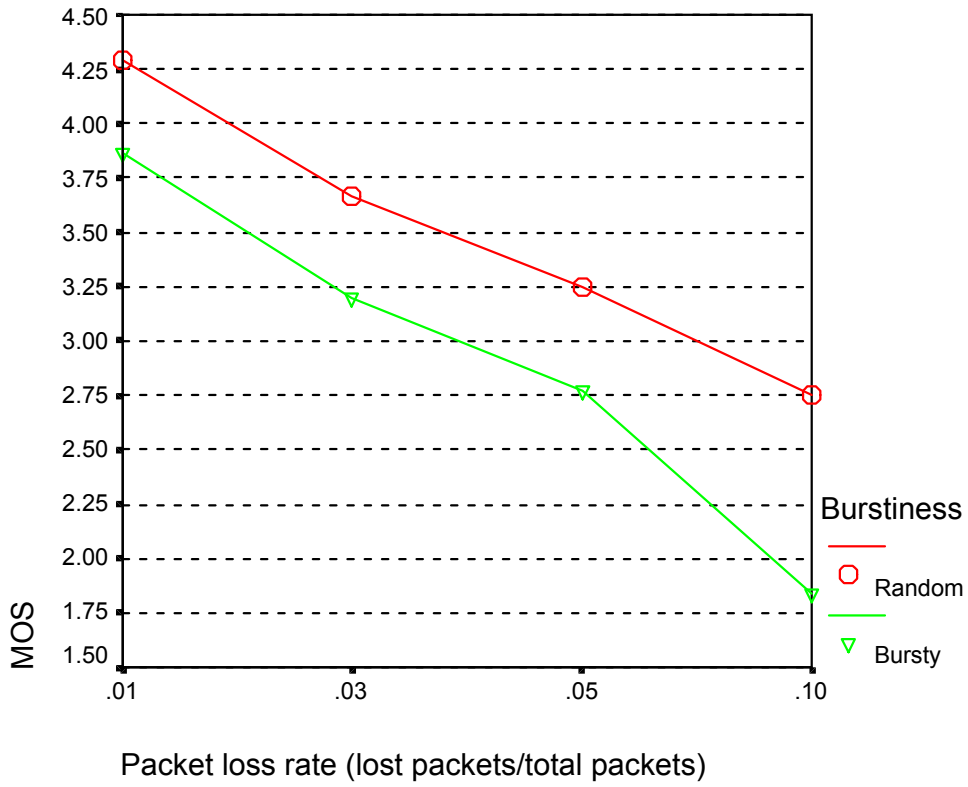

Figure 2. MOS scores by "burstiness" and packet loss rate

ITU-T PERFORMANCE DATA

The competition designed by the ITU-T used two sets of data. The first set consisted of 22 benchmark speech databases used by developers during the creation of candidate algorithms. These speech databases contained a variety of impairments intended to be representative of existing and emerging telecommunications networks. The second set consisted of 8 additional speech databases that were unknown during the development of PESQ. The performance data from the recent ITU-T competition showed that PESQ significantly outperformed all of the other candidate algorithms. However, only PESQ and PSQM data is presented here.

It is important to note that all of ITU-T performance measures were calculated after a 3rd order polynomial regression mapping was performed for each of the speech databases separately (as described in the previous section).

Correlation analysis

Table 1 shows the correlations (r) between MOS and PESQ scores for several data sets. Below is a brief description of each of the data sets:

- **ITU-T known**—This data set is made up of the 22 benchmark databases that were used in the validation of the new ITU-T recommended method for objectively measuring speech quality (ITU-T Recommendation P.862). It contains a wide range of conditions intended to be representative of existing and emerging telecommunications networks.

- ***PSQM (ITU-T known)*—**This shows performance of the PSQM algorithm on the same 22 benchmark databases described above in the "ITU-T known" data set.
- **ITU-T unknown**—This data set consists of 8 speech databases that were unknown during the development of the PESQ algorithm. Therefore, they can be viewed as a validation data set. Two of these speech databases came from Lucent's speech coding group and are further analyzed in this paper.
- **MPAC data by file**—This speech database comes from Lucent's Multimedia Perception Assessment Center (MPAC). It consists of speech samples that were subjected to various packet loss rates, burstiness of packet loss, Packet Loss Concealment (PLC) strategies, IP payload size, and speech coders.
- **MPAC data by condition**—This is the same speech database as "MPAC data by file", except that both PESQ and MOS scores were averaged per condition before calculating accuracy measures.

It can be seen from Table 1 that correlations for all of the speech databases are high. However, this does not necessarily mean high accuracy. Correlations ignore "absolute differences" and "differences in scale" between objective and subjective scores.

Also, correlations seem more impressive than they really are [4]. A correlation of .95 only reduces the standard error of estimates (i.e., prediction error) by two-thirds. Table 2 shows how the standard error of estimates decreases as the correlation increases.

Data Set	RMSE	r	r^2	Absolute Error		
				Mean	<0.25	<0.5
ITU-T known	--	.935	.874	--	69%	91%
PSQM (ITU-T known)	--	*.809*	*.654*	--	*47%*	*76%*
ITU-T unknown	--	.930	.865	--	72%	91%
MPAC data by file	.50	.921	.849	0.41	34%	62%
MPAC data by condition	.45	.955	.912	0.38	38%	69%

Table 1. Overall indicators of PESQ accuracy.

NOTE 1: *The data set labeled "PSQM (ITU-T known)" indicates accuracy of the PSQM algorithm and is included for comparison.*

NOTE 2: *The "ITU-T known", "PSQM (ITU-T known)", and "ITU-T unknown" used regression mapping on a per-study basis before calculating performance measures. This will result in overly optimistic performance estimates.*

Correlation (r)	Standard error of estimates (s_{Y-X})
0	s_Y
.50	$0.886s_Y$
.80	$0.600s_Y$
.866	$0.500s_Y$
.90	$0.436s_Y$
.95	$0.312s_Y$

Table 2. Decrease in standard error of estimates as correlation increases

NOTE 1: "s_Y" *represents variability is subjective ratings.*

Errors in prediction

Table 1 also shows statistics on distributions of prediction errors. For example, The "MPAC data by file" data set indicates that the average distance between the predicted MOS and observed MOS was 0.41. Also, the distance was less than ±0.25 34% of the time and ±0.5 62% of the time. An important difference between the Lucent data and ITU-T data is that for the ITU-T data, PESQ scores were transformed using a monotonically constrained 3rd order polynomial before calculating the residual error. This is why the ITU-T performance data is so much better than the Lucent data. As described in the previous section, regression mapping on a per study basis has the effect of removing prediction errors that the PESQ algorithm should be penalized for. Therefore, the prediction error data presented by the ITU-T suggest higher accuracy than what can be expected by the telecommunications industry.

LUCENT PERFORMANCE DATA

There are three different Lucent subjective studies. One study was conducted by Lucent's Multimedia Perception Assessment Center (MPAC). Experimental variables in this study were speech coder, packet loss concealment strategy, IP payload size, packet loss rate, and burstiness of packet losses. The other two studies were conducted by Lucent's speech coding group. It is also worth noting that these two studies were part of the "ITU-T unknown" data set presented in the previous section. Most of the analysis that follows was performed on the MPAC data set. Whenever data from the speech coding group is presented it will be explicitly stated.

The MPAC subjective study, and calculation of PESQ scores, were carried out in accordance with ITU-T Recommendations P.830 [5] and P.862 [1]. There were 50 participants in the MPAC subjective study. Participants listened to sound recordings over a telephone handset that had the appropriate receive frequency response.

Results were analyzed on a per file basis instead of a per condition basis. This was done because it is likely that many users of the algorithm will make decisions based on per file performance data and because this is also likely to represent worst case performance for the PESQ algorithm (there is a tendency for performance to improve when condition averaging is performed). Furthermore, ITU-T Recommendation P.862 does not indicate that calculating performance on a per file basis is inappropriate.

Predicting end-user acceptance of transmission quality

One of the ways the PESQ algorithm will be used is to predict the Mean Opinion Score (MOS) of quality as judged from a panel of end-users. To determine the accuracy of the PESQ algorithm in answering this question, the residual errors (i.e., prediction errors) must be analyzed. The residual error is simply the difference between the predicted MOS and the MOS obtained through perceptual testing. For example, if the PESQ score is 3.7, but the observed MOS is 4.1, then the residual error is –0.4 (3.7-4.1=-0.4). Also, the *absolute value* of the residual error, or absolute error, is 0.4 (|-0.4|=0.4).

Looking at a scatterplot is a useful way to identify the nature of prediction errors. Figure 3 shows a scatterplot of MOS and PESQ scores. Each point on the plot represents a test case, and has an associated MOS and PESQ score. For each data point on the graph, the difference between the x-axis variable (i.e., MOS) and y-axis variable (i.e., PESQ score) represents the prediction error for that particular speech file. If MOS were perfectly predicted by PESQ, then the scores would fall along a straight line with a slope of 1 that goes through the origin. Such a reference line is provided in Figure 3.

It can be seen that there is some agreement between MOS and PESQ scores. In general, as perceived quality goes down so do PESQ scores. However, the scatterplot also shows a couple of accuracy issues. First, there is a tendency for PESQ to underpredict performance when perceived quality is high, and overpredict performance when perceived quality is low. Second, the algorithm is not accurate enough to reliably detect small differences in performance. If it were, then one would expect to see the range in PESQ scores, for a given MOS score, to be smaller. For example, when the MOS was 3.0, PESQ scores ranged from 2.6 to 3.6. This is a very large amount of variability—especially considering this is a 5-point scale and the range of scores is closer to 4 (i.e., 1.5 to 4.5).

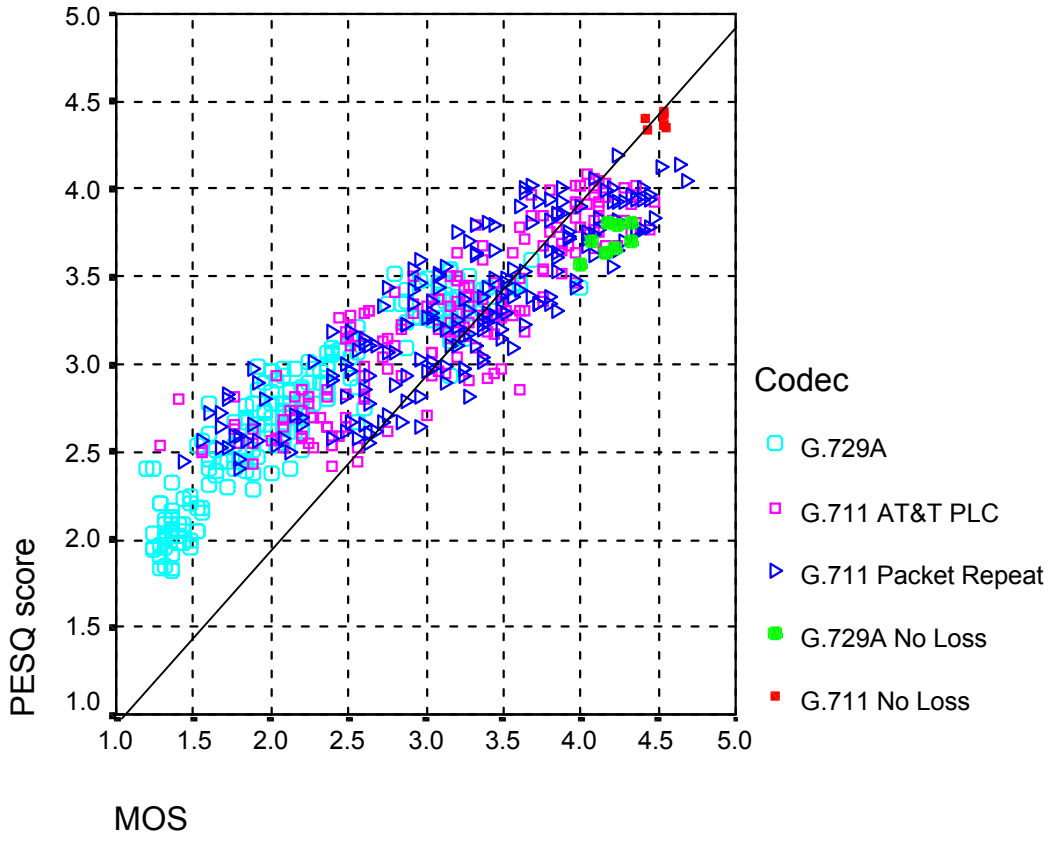

Figure 3. Scatterplot of MOS and PESQ scores

The same information presented in Figure 3 can also be displayed in different ways that help visualize the data. Figure 4 shows the distribution of residual errors for the MPAC data. Overall statistics of this distribution were given in Table 1. Note that the mean value shown in the Legend of Figure 4 is different than that shown in Table 1. This is because residual errors are shown in Figure 4 and the *absolute value* of residual errors is summarized in Table 1.

If PESQ perfectly predicted perceived quality then the residual errors would all be clumped around 0. However, it can be seen that PESQ scores often differed from MOS scores by more than ±0.25, and overpredicted performance more often than underpredicted performance.

Figure 5 shows residual errors by MOS. This is an intuitive way of looking at the magnitude of prediction errors as a function of MOS. If PESQ perfectly predicted MOS, then all of the data points would fall on a horizontal line with a value of 0.00. Note that although the prediction errors are small numbers (i.e., often less than 1), they are a significant portion of the usable range of the MOS scale.

The prediction errors in Figure 5 can be summarized by using Boxplots. Boxplots are useful in visualizing distributions of scores, and will be used extensively in the following sections. Figure 6 shows boxplots of residual errors by MOS. The thick black line within each box represents the median value of residual error (i.e., 50th percentile). 50% of the scores fall above this value, and 50% fall below it. The top of the box represents the 75th percentile. 25% of the scores fall above this value and 75% fall below it. The bottom of the box represents the 25th percentile. 75% of the scores fall above this value and 25 % fall below it. The "whiskers" (i.e., lines extending out away from the box) will extend to include the full range of values up to 1.5 times the length of the box. Values beyond 1.5 times the length of the box are considered "outliers" and are indicated with the symbol "o". Values more than 3 times the length of the box are considered "extreme" scores and are indicated with the symbol "*". Text labels for outliers and extreme scores can also be seen on the graphs. They provide descriptive information about these cases. "N=" near the origin of the graph refers to the number of observations represented by each boxplot, and these values are shown below each boxplot along the horizontal axis.

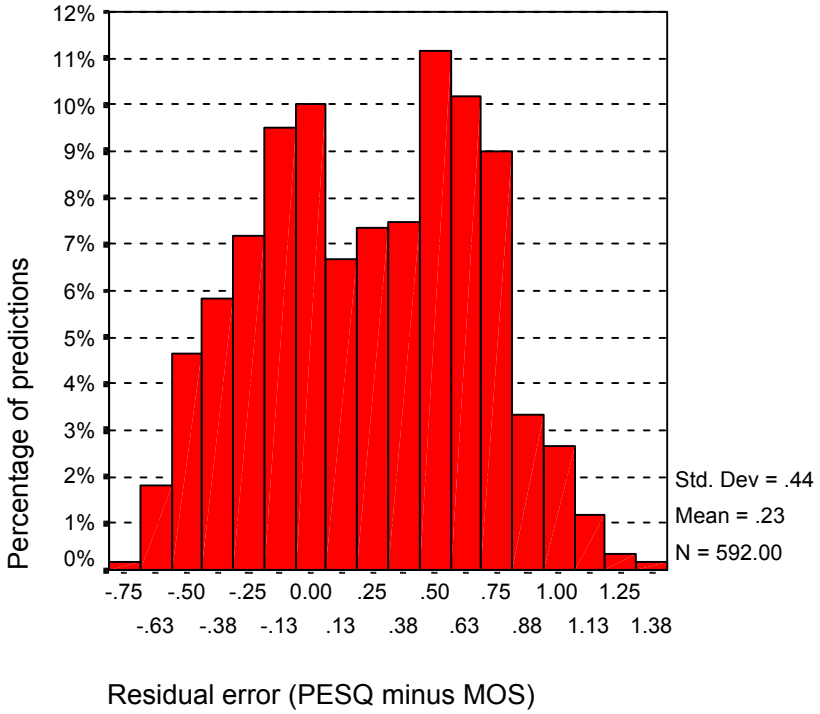

Figure 4. Distribution of residual errors

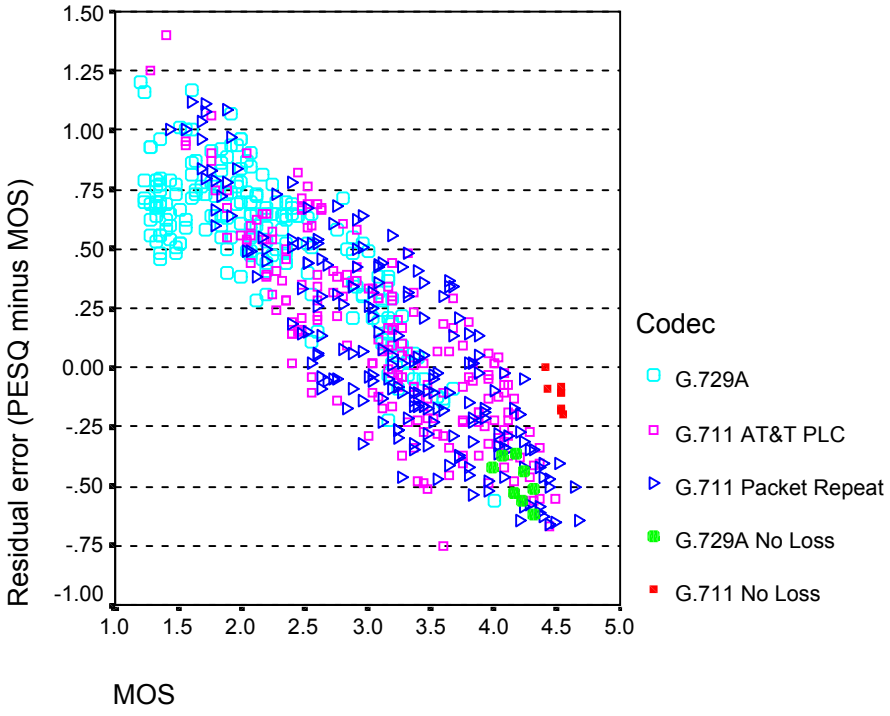

Figure 5. Scatterplot of residual errors by MOS

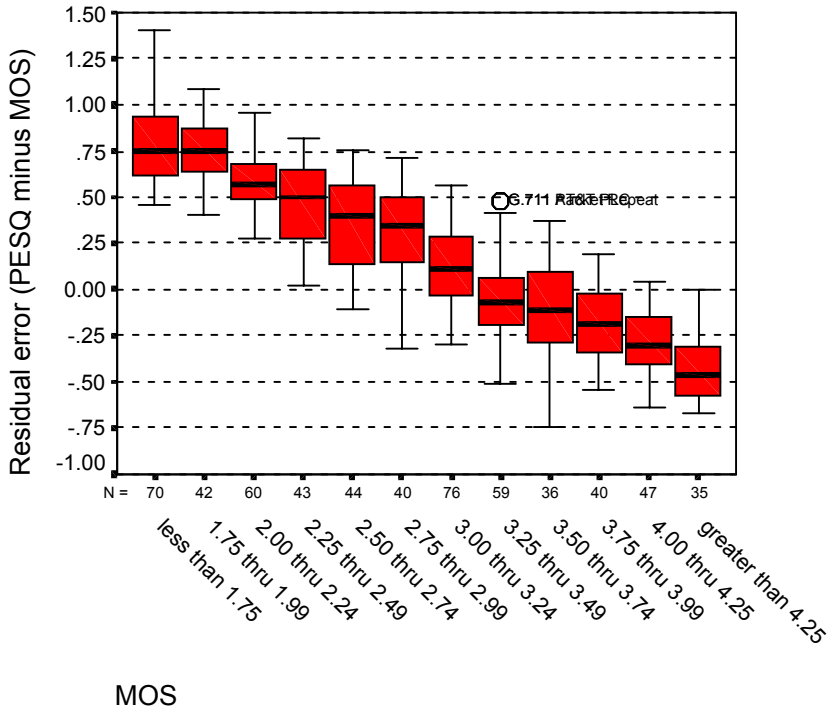

Figure 6. Boxplots of residual errors by MOS

It can be seen in Figure 6 that the errors in prediction depended on the perceived quality. When perceived quality was high there was a tendency to underpredict performance, and when perceived quality was low the performance was overpredicted.

Figure 7 shows distributions of prediction errors for various ranges of PESQ scores. This graph should be interpreted with caution, however, because the results (i.e., errors) of this study may not generalize to other data sets.

Figure 8 shows the probability that the MOS is within ±0.25 of the PESQ score as a function of PESQ. This graph highlights the fact that accuracy decreases as PESQ scores decrease.

Accuracy of PESQ depended on the system under test

A statistical analysis was performed to assess the relationship between PESQ accuracy and the following experimental variables: Speech coder (using various error concealment schemes), packet loss rate, IP payload size, and "burstiness" of packet losses. A log-linear statistical analysis was performed. Log-linear statistical techniques use expected cell frequencies to assess relationships among experimental variables. Log-linear techniques were chosen because they do not require the assumptions of normality and homogeneity of variance that other parametric statistical approaches do (e.g., repeated measures ANOVA). Since log-linear analysis uses categorical variables, the residual error was transformed from a continuous variable into a categorical variable with two levels. The first level of this new variable consisted of those cases where the absolute error was less than ±0.25. All other cases were assigned to the other level. Therefore, accuracy was represented by the proportion of cases falling into these two levels.

The statistical analysis showed that PESQ accuracy depended on speech coder, packet loss rate, IP payload size, and "burstiness" of packet losses. There were 3 significant 2-way interactions, and 3 significant main effects.

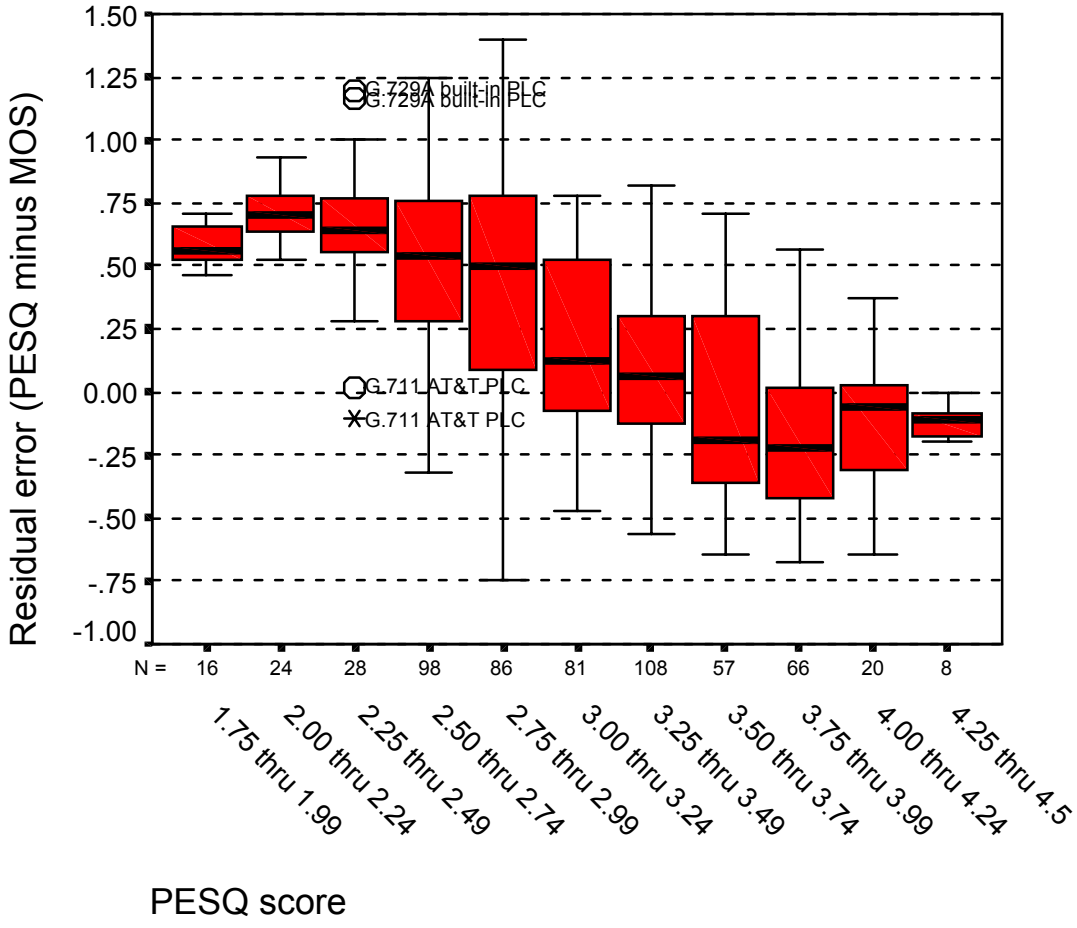

Figure 7. Distributions of residual errors by PESQ

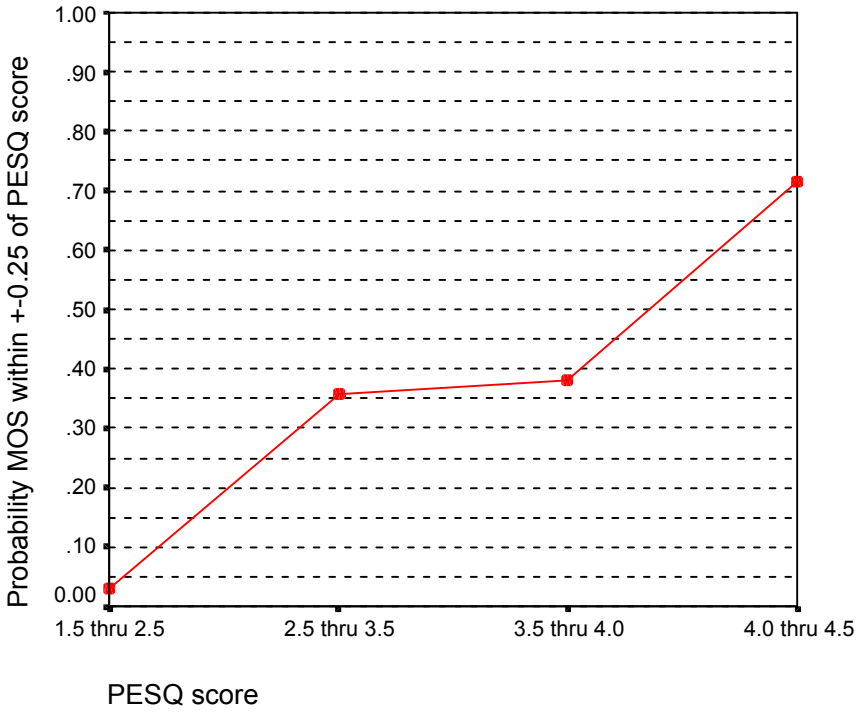

Figure 8. Probability that MOS within ±0.25 of PESQ score by PESQ

There was a significant LOSS RATE X SPEECH CODER interaction (p<.001). This interaction can be seen in Figure 9. It can be seen that the effects of loss rate depended on the speech coder. For G.711 with both types of PLC, PESQ accuracy gradually decreases with increasing loss rate. For G.729A, however, there is a big decrease in PESQ accuracy going from 1% to 3% packet loss, and very little change in accuracy from 3% up to 10% packet loss.

There was also a significant LOSS RATE X BURSTINESS interaction (p<.001). This interaction can be seen in Figure 10. When the packet losses were random, PESQ accuracy slowly decreases with increasing packet loss rate. When the packet losses were "bursty" (i.e., occur close to each other), then PESQ accuracy decreased rapidly with increasing packet loss rate. It is also worth noting that perceived quality also falls off more rapidly with "bursty" packet loss than with random packet loss; suggesting this result is due to an inability of PESQ to accurately track this perceptual effect. Note that the absolute value of the residual error is plotted instead of the residual error itself. This was done to make the interaction more obvious.

The last significant 2-way interaction was the SPEECH CODER X BURSTINESS interaction (p<.001). This interaction can be seen in Figure 11. Figure 11 shows that the affect of burstiness on PESQ accuracy is more pronounced for the G.711 Packet Repeat speech coder than with the other speech coders.

Figure 12 shows a significant main effect of speech coder (p<.001). It can be seen that PESQ accuracy was strongly dependent on system under test. Accuracy was similar for G.711 between the 2 types of packet loss concealment. However, performance of G.729A with built-in error concealment was consistently overpredicted, and there was less variability in the residuals. It is also worth noting that prediction errors were different for G.711 and G.729A under "no loss" conditions than they were when there was packet loss for these codecs.

Figure 13 shows a significant main effect of packet loss rate (p<.001). It can be seen that there is a systematic affect of packet loss rate on PESQ accuracy.

Figure 14 shows the affect of IP payload size on PESQ accuracy. As IP payload size increases, accuracy gets worse.

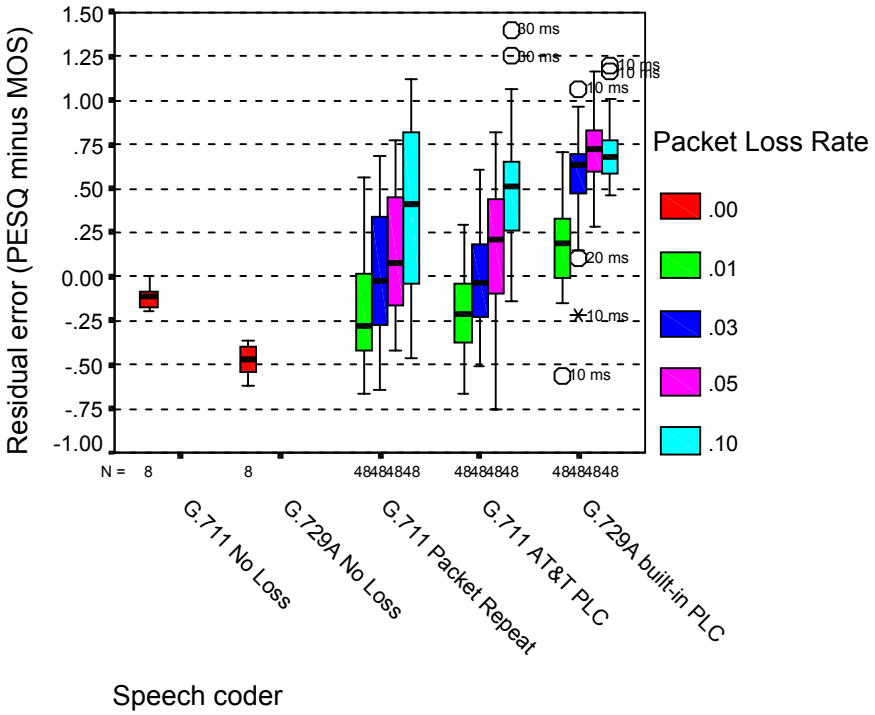

Figure 9. Affects of LOSS RATE and SPEECH CODER on PESQ accuracy

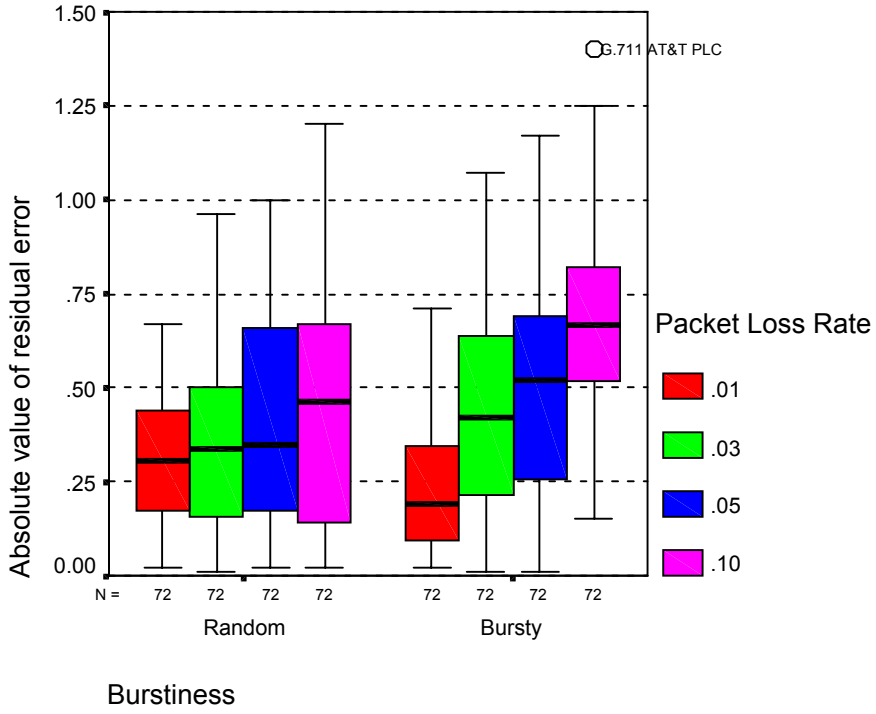

Figure 10. Affects of LOSS RATE and BURSTINESS on PESQ accuracy

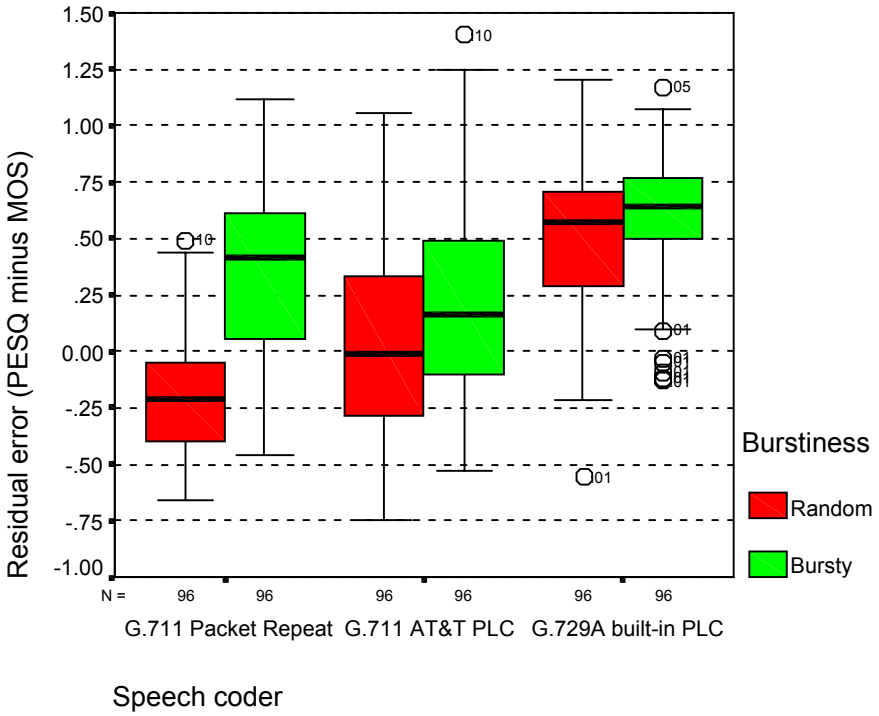

Figure 11. Affects of SPEECH CODER and BURSTINESS on PESQ accuracy

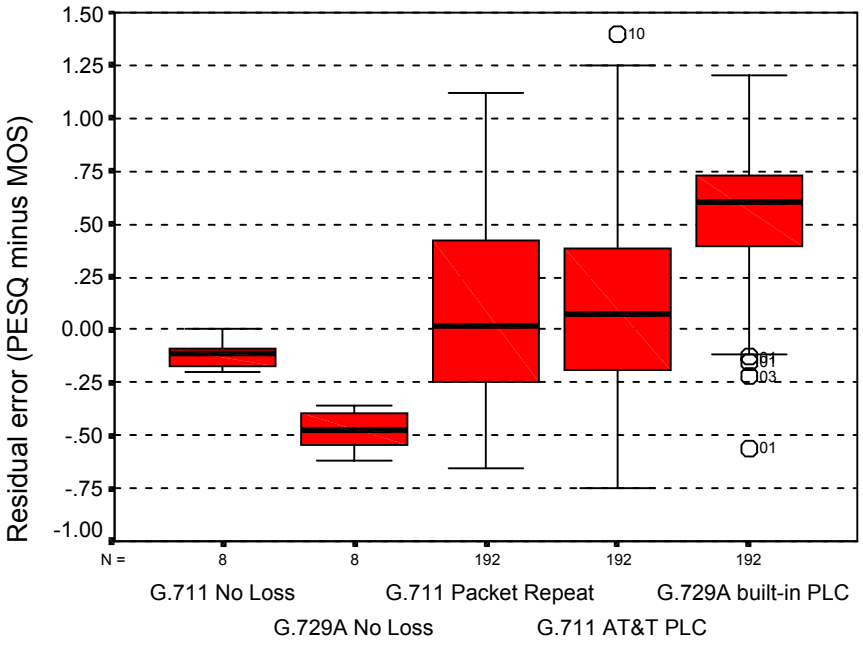

Figure 12. Affect of speech coder on PESQ accuracy

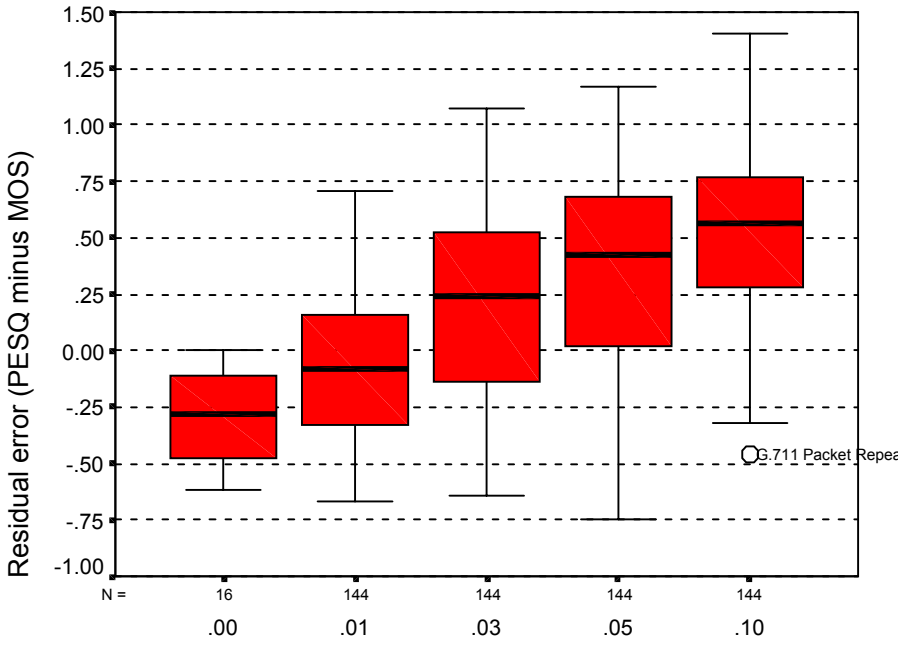

Figure 13. Affect of packet loss rate on PESQ accuracy

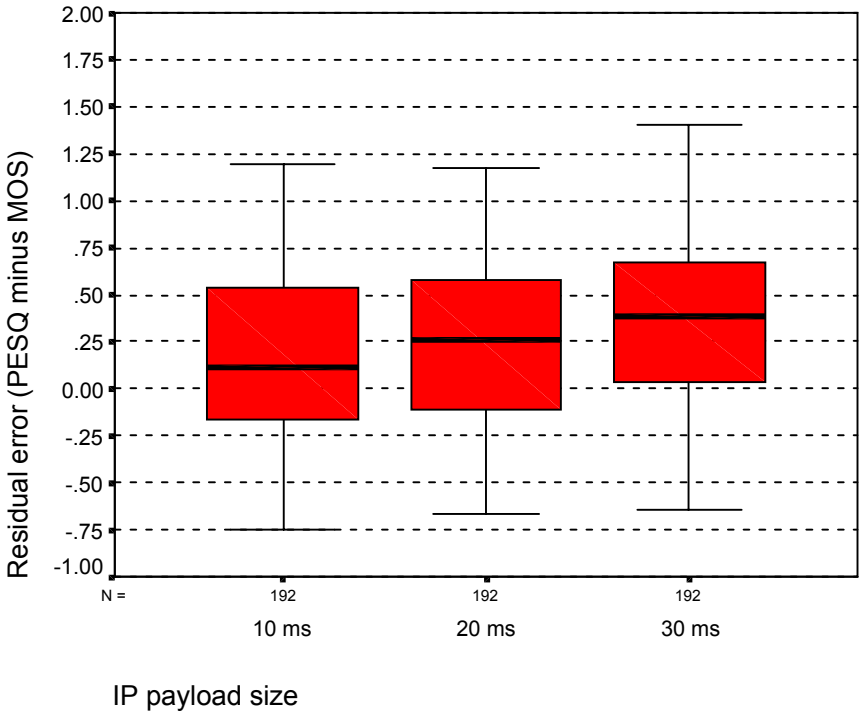

Figure 14. Affect of IP payload size on PESQ accuracy

Validating voice quality performance using pass/fail criterion

One of the ways PESQ is being used is to determine whether a system under test will provide acceptable transmission quality. This is usually done by setting a performance criterion and passing those systems that exceed the threshold while failing the systems that fall below the threshold. A "toll quality" criterion value of 4.0 is known to be used by some in the telecommunications industry. Therefore, this was used in the current analysis to determine the accuracy of PESQ as a method of validating voice quality using a pass/fail criterion.

Table 3 shows the accuracy of the PESQ algorithm in predicting whether quality would be judged better than 4.0 by participants. When participants judged quality above 4.0, PESQ correctly identified these cases only 26% of the time. In other words, PESQ incorrectly classified these cases 74% of the time. The practical significance of this is that PESQ will often indicate that perceived quality is not good enough when in fact it is.

It is possible that some of the classification errors resulted from measurement error in the subjective data. In other words, some of the MOS cases categorized above 4.0 could be due to measurement error in subjective scores instead of them truly being above 4.0. To address this concern, another analysis of categorization errors was performed where only those cases that had a high degree of confidence of being above 4.0 were analyzed. A high degree of confidence was defined as a statistically significant result from a one-tailed t-test that the population mean for the sampled data was above 4.0 (alpha=.05). Even under this constraint, PESQ still incorrectly categorized these cases as being below 4.0 67.5% of the time. Note that the correlation coefficient for this data is .92! This is another example of how the ITU-T data suggest higher accuracy than what can be expected when PESQ is used by the telecommunications industry.

			PESQ score		
			Below 4.0	Above 4.0	Total
MOS	Above 4.0	Number of Conditions	57	20	77
		% of Row Total	74.0%	26.0%	100.0%
	Below 4.0	Number of Conditions	509	6	515
		% of Row Total	98.8%	1.2%	100.0%
Total		Number of Conditions	566	26	592
		% of Row Total	95.6%	4.4%	100.0%

Table 3. PESQ accuracy when using 4.0 as pass/fail criterion

It is also possible that the results from the current study are not typical of PESQ performance. The quality of the particular speech files used or the actual participants in the study could have affected the MOS scores given, and thus affected performance of PESQ. However, this does not seem to be the case. Two other speech databases from Lucent's speech coding group were analyzed and similar results were found. It is also worth noting that these made up 2 of the 8 "ITU-T unknown" speech databases used in the validation of PESQ by the ITU-T. Therefore, it would be hard to argue that there are problems with these speech databases as well since they were used as evidence to the validity of the PESQ algorithm. PESQ performance for the LUWI96 (waveform interpolation coder) and LURC98 (RCELP coder) speech databases are shown in Tables 4 and 5, respectively.

Again, it can be seen that PESQ often indicates signal quality is not good enough when in fact it is.

			PESQ score		
			Below 4.0	Above 4.0	Total
MOS	Above 4.0	Number of Files	20	16	36
		% of Row Total	55.6%	44.4%	100.0%
	Below 4.0	Number of Files	83	1	84
		% of Row Total	98.8%	1.2%	100.0%
Total		Number of Files	103	17	120
		% of Row Total	85.8%	14.2%	100.0%

Table 4. PESQ accuracy for LUWI96 data set

			PESQ score		
			Below 4.0	Above 4.0	Total
MOS	Above 4.0	Number of Files	69	45	114
		% of Row Total	60.5%	39.5%	100.0%
	Below 4.0	Number of Files	338	28	366
		% of Row Total	92.3%	7.7%	100.0%
Total		Number of Files	407	73	480
		% of Row Total	84.8%	15.2%	100.0%

Error rates

Table 5. PESQ accuracy for LURC98 data set

Specifying voice quality performance in a contract (e.g., SLA)

Current results indicate that PESQ often indicates objectionable signal quality when in fact it is judged "Good". Furthermore, it will consistently give the wrong answer for some network connections. This problem cannot be overcome by statistics, as some have suggested.

Objective algorithms, such as PESQ, are inappropriate for legal contracts specifying performance as long as there is a reasonable chance that failing to meet the requirement can be caused by a problem with the objective algorithm instead of a problem with the performance of the system under test.

Using PESQ for competitive analysis

PESQ can be used to compare performance among competing algorithms or telecommunications devices. It can also be used in system optimization (i.e., parameter "tweaking"). To assess accuracy of PESQ for these applications, pairs of test conditions were compared. For each paired comparison, the difference in PESQ scores between the two test items was compared to the difference in MOS scores between them. There were four possible outcomes to these comparisons:

- **Correctly identified**—PESQ correctly identified the relationship between the two test items (i.e., A better than B, B better than A, or no difference).
- **Failed to detect difference**—PESQ indicated no significant difference when the participants consistently preferred one test item to the other.
- **Falsely detected difference**—PESQ indicated that one test item was better than the other, when participants found no preference for one test item over the other.
- **Opposite conclusion**—Participants consistently preferred one test item to the other, but PESQ indicated that the less preferred test item was better. For example, if participants preferred A over B, then PESQ indicated that B was better than A. This is the most interesting category for assessing performance of PESQ, since it represents those cases where PESQ scores led to the wrong conclusion.

A matched-samples t-test was performed for each comparison to determine if the difference between MOS scores was significant. An alpha=.05 was used in these statistical tests.

The same criteria for testing differences between MOS scores were applied to PESQ scores. For example, if the analysis of the subjective data indicated that the distance between the MOS scores had to be 0.20 in order to be significant, then the difference between PESQ scores also had to be 0.20 in order to be considered significantly different. It was felt that this was reasonable given that PESQ scores are supposed to be using the same scale as MOS scores, and that a method was needed to distinguish small, insignificant

differences, from larger differences indicating one item was better than the other.

It should also be mentioned that that a paired-comparison was classified as an "opposite conclusion" error if there was a statistically significant preference for one item, but PESQ indicated a preference in the opposite direction—regardless of the magnitude. This seemed like the appropriate thing to do given that for two systems where all other things are equal, the system with the higher PESQ score will be selected. For example, if system A has a PESQ score of 4.10 and system B has a score of 4.11, then system B will be selected when all other things are equal. In this case, if end-users actually preferred system A, then PESQ would have led to the wrong conclusion.

Table 6 shows the frequency of each of the possible outcomes for all the within-subjects paired comparisons. It can be seen that PESQ correctly identifies end-user preferences most of the time (70%). PESQ suggests that there are no noticeable differences in performance when participants have a preference around 15% of the time. PESQ suggests that there is a noticeable difference in performance when participants don't have a consistent preference around 10% of the time. Finally, PESQ leads to the wrong conclusion around 5% of the time.

Although "opposite conclusion" errors only make up 5% of the paired-comparisons in Table 6, the probability for this type of error occurring when using PESQ for competitive analysis and system optimization may be higher. This is because these types of errors are more likely when the test items are of similar quality. Table 6 contains all paired-comparisons, including those that are at different ends of the quality spectrum and very easy to distinguish. For example, comparing G.711 under no packet loss (i.e., very high quality) with G.711 under 10% packet loss (i.e., very low quality) is very easy for any objective measure to accurately predict and is not likely to represent test items included in a competitive shoot-out or system optimization.

To more accurately assess the probability of opposite conclusion errors in real-world use, further analysis was done on only those paired-comparisons where both test items had MOS scores above 3.5. Figure 15 shows the probability of opposite conclusion errors for the subset of paired-comparisons where there was a statistically significant preference for one item over the other by participants, and MOS scores for both items were above 3.5. The probability of an opposite conclusion error is shown as a function of the difference in PESQ scores between the two test items. It can be seen in Figure 15 that for PESQ differences below around .25, the chances of selecting the wrong item start to get large. It should be pointed out that even when the probability of an opposite conclusion error is .40, this is still better than randomly selecting one of the two test items (randomly picking would result in a value of .50). However, the options for competitive analysis are not to use PESQ or randomly pick. The current data suggest that these decisions should still be based on perceptual testing.

It should also be noted that the chances of an opposite conclusion error will probably depend on the nature of the comparisons. If the nature of the impairments introduced are uni-dimensional in nature (e.g., tweaking a single parameter), then opposite conclusion errors will probably be less likely than if they are multi-dimensional in nature (e.g., comparing different technologies with different impairment sets).

To summarize, the current results indicate that there are limitations to using PESQ for competitive analysis. While PESQ can consistently distinguish large differences in perceived performance, there are risks associated with using it to select among competing systems that are fairly close in quality.

Outcome	# of comparisons	%
Correctly identified	4,291	70%
Failed to detect difference	911	15%
Falsely detected difference	641	10%
Opposite conclusion	323	5%
Total	*6,166*	*100%*

Table 6. Paired-comparison outcomes

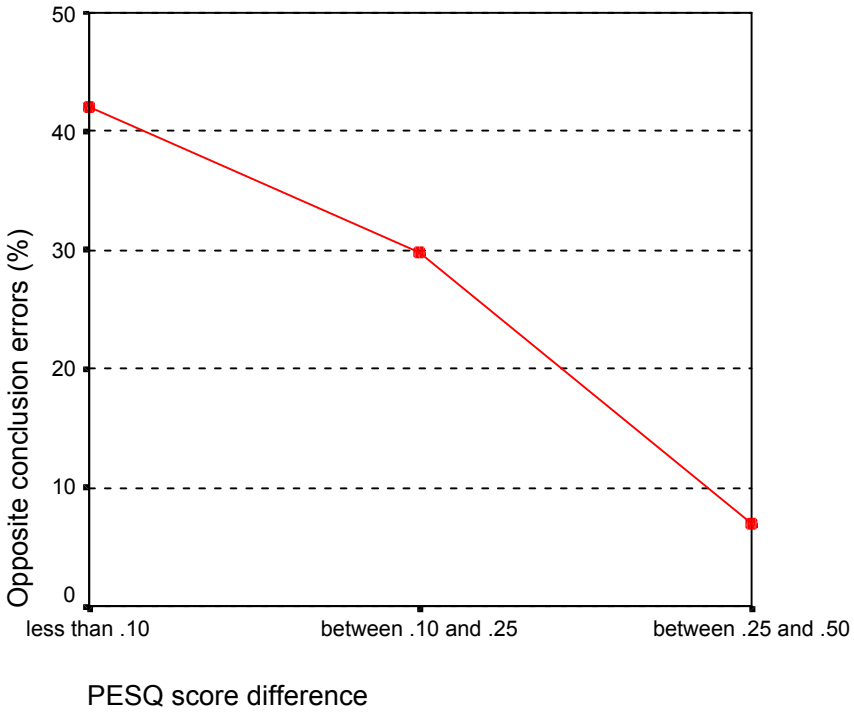

Figure 15. Probability of opposite conclusion error by difference in PESQ scores

CONCLUSIONS

The ITU-T data suggest higher accuracy than what can be expected by the telecommunications industry in real-world use. There are a couple of reasons for this.

First, the high correlations presented in ITU-T Recommendation P.862 seem to imply that PESQ will be very accurate—especially to the reader who may not be knowledgeable in statistics. However, correlations are not measures of accuracy, they are measures of covariance. Correlations can be high even for measures that are not very accurate. For example, even when the correlation coefficient was .92, 74% of network connections that were judged higher than 4.0 by listeners were incorrectly classified by PESQ as being below 4.0. Also, the chances of selecting the wrong network connection (i.e., test item) in a competitive analysis got as high as 40% when test items were close in quality.

Second, prediction errors are only calculated after data fitting PESQ scores to the results of each subjective study. As described earlier, this has the effect of reducing prediction errors that truly exist. Therefore, the residual errors presented in ITU-T Recommendation P.862 indicate higher accuracy than what will be experienced from real-world use by the telecommunications industry.

Accuracy of PESQ depended on the nature of the network connection. Speech coder (w/ various PLC), packet loss rate, IP payload size, and “burstiness” of packet losses all affected how accurate PESQ was. The practical significance of this finding is that the currently available performance data may not generalize to a particular network connection being evaluated. Also, PESQ will consistently give inaccurate results for some systems. This deficiency cannot be overcome by statistics (e.g., repeatedly measuring performance and using a statistic such as the mean to represent performance).

The current results indicate that there are problems with using PESQ as a method of verifying speech quality. PESQ often indicated that speech quality was “objectionable” when in fact listeners judged it as

"good". This result was found in three independent studies, two of which were used by the ITU-T in the validation of PESQ. PESQ should be viewed as an imperfect predictor of speech quality—not the final word. If PESQ indicates degraded speech quality, this should be verified by listening to the system under test and/or more formal perceptual studies.

There are also limitations to using PESQ in competitive analysis. PESQ often led to the wrong conclusion when comparing performance of networks with similar quality. It is worth noting, however, that PESQ may perform better when the impairments are uni-dimensional in nature (e.g., only differ in level of noise) than when they are multi-dimensional.

PESQ is not accurate enough for use in legal contracts specifying speech quality requirements (e.g., SLA). This is the case whenever there is a reasonable chance that failing to meet the agreement can be caused by a problem with the measurement method, instead of a problem with network performance.

To summarize, while PESQ can be a useful tool in helping identify potential problem areas, there are limitations to using PESQ for verification of speech quality performance, competitive analysis, and system optimization. Also, PESQ is not accurate enough to specify speech quality requirements. Important decisions should still be based on the results from well-designed subjective studies.

REFERENCES

[1] ITU-T Recommendation P.862 (02/2001), Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs

[2] ITU-T COM12-D109 (09/1999), Subjective evaluation of packet loss concealment (PLC) techniques, Nortel Networks

[3] ITU-T COM12-D114 (09/1999), Results of a subjective listening test for G.711 with frame erasure concealment, AT&T

[4] D.C. Howell, Statistical methods for psychology, Boston MA, PWS-KENT, 1992, ch. 9, pgs. 239-240.

[5] ITU-T Recommendation P.830 (02/1996), Subjective performance assessment of telephone-band and wideband digital codecs

[6] ITU-T COM12-D136 (05/2000), Performance of the integrated KPN/BT objective speech quality assessment model, KPN Research and BT